EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Selecting scenes for 2D and 3D subjective video quality tests

Margaret H Pinson[1*†], Marcus Barkowsky[2†] and Patrick Le Callet[2]

Abstract

This paper presents recommended techniques for choosing video sequences for subjective experiments. Subjective video quality assessment is a well-understood field, yet scene selection is often driven by convenience or content availability. Three-dimensional testing is a newer field that requires new considerations for scene selection. The impact of experiment design on best practices for scene selection will also be considered. A semi-automatic selection process for content sets for subjective experiments will be proposed.

Keywords: Video quality; Subjective testing; Stereoscopic 3D; Content selection; Video quality assessment; Scene selection

Introduction

Advances in the consumer video market include higher resolution (Full-HD, 4K, 8K), additional dimensions and reconstruction precision (three-dimensional (3D), high dynamic range, wide color gamut), associated media (audio wavefield synthesis, emotive devices), and interaction methods (mobile phones, social networks, see-through glasses with augmented reality). These technologies have caused a renewed interest in the analysis of quality of experience through subjective and objective measurement methods. This rapid development in audiovisual technology creates new challenges in quality of experience assessment and analysis, and these new challenges require new subjective experiments to be conducted. In most cases, subjective experiments are necessary to establish ground truth data that helps in training, verification, and validation of objective measurement methods. In order for these subjective experiments to add value to the research community, testing conditions must be carefully considered.

Regardless of future audiovisual technology, scene selection will always be one important component of the testing conditions for video quality testing. Selection should be based on video characteristics and the purpose of the experiment, not on personal preference or convenience. For emerging video technologies, limited content choice may be a major factor. It is therefore important to provide guidelines that identify content that is suitable for testing the emerging technologies.

In this paper, we first describe guidelines for scene selection for traditional, entertainment-oriented tests that were developed using over two decades of experience in designing video quality subjective tests. Second, we explore new issues that are critical for selecting 3D scene content. Third, we review experimental design considerations common to both types of subjective testing.

Basic scene selection

Entertainment-oriented subjective video quality tests try to represent a wide range of entertainment content in a scene pool containing approximately eight to ten clips. It is impossible to include every genre and visual effect with only eight or ten clips, but approaching this ideal improves a test's accuracy.

Avoid offensive content

It is important to avoid subject matter that may be offensive, controversial, polarizing, or distracting:

- Violence
- Indecent, revealing, or suggestive outfits
- Erotic situations
- Drugs and drug paraphernalia
- Politics

* Correspondence: mpinson@its.bldrdoc.gov
†Equal contributors
[1]National Telecommunications and Information Administration (NTIA), Institute for Telecommunication Sciences (ITS), U.S. Department of Commerce (DOC), 325 Broadway St, Boulder, CO 80305, USA
Full list of author information is available at the end of the article

- Religion
- Horror films
- Medical operations

Offensive, controversial, polarizing, or distracting content may change how subjects rate the video sequences. Subjects may give the clip a lower mean opinion score (MOS) or fail to pay close attention to the rating task.

Consider content editing and camerawork

The impact of scene content editing and camerawork cannot be underestimated. Viewer instructions for subjective testing should include a statement such as: 'Please do not base your opinion on the content of the scene or the quality of the acting.' Yet ratings inevitably include both the clip's artistic and technical qualities.

To illustrate this issue, we will examine the scenes selected for the Video Quality Experts Group (VQEG) high-definition television test [1]. This international test produced six subjectively rated databases. The six scene pools were carefully selected by Margaret Pinson to have very similar objective characteristics. During the selection process, all original scenes were judged to have a quality of 'good' or better by an expert panel of video quality subjective testing researchers. Each dataset included 13 original video sequences, which were rated on the absolute category rating (ACR) 5-level scale of excellent = 5, good = 4, fair = 3, poor = 2, and bad = 1. The ratings from all subjects were averaged to compute a MOS. These MOS were rank-sorted to identify the original video sequences in each dataset that have the highest and lowest MOS. Figures 1 and 2 show sample frames from the original sequences with the highest and lowest MOS, respectively.

The average MOS drops from 4.7 in Figure 1 to 4.1 in Figure 2. So on average, the available range for subjects' ratings of the impaired video sequences shrank:

- On average by 16% (i.e., from (4.7 to 1) to (4.1 to 1))
- At most by 28% (i.e., dataset vqegHD6, from (4.9 to 1) to (3.8 to 1))
- Theoretically up to half of the scale (i.e., from (5.0 to 1) to (3.0 to 1)) if sequences with 'fair' quality had been allowed

The precision of subjective ratings is more a trait of the subject than the scale, as demonstrated by Tominaga et al. [2]. The distribution of ratings will not narrow simply because subjects have a smaller portion of the scale to work with when rating all versions of some sequences. In practical terms, this means that the data analysis will be less able to distinguish between distortions for sequences with poor editing and camerawork.

The impact of editing and camerawork can be seen in these sequences. Figure 1 scenes contain more scene cuts, animation, vibrant colors, and good scene composition. These qualities add visual interest and improve the esthetic appeal. Figure 2 sequences contain a variety of minor problems that had a large cumulative impact on MOS, such as motion blur, analog noise, camera wobble, poor scene composition, long shot lengths (e.g., no scene cuts), a boring topic, or an uninteresting presentation (e.g., the action in vqegHD5 src2 would be more exciting if seen from a closer zoom). These minor problems have

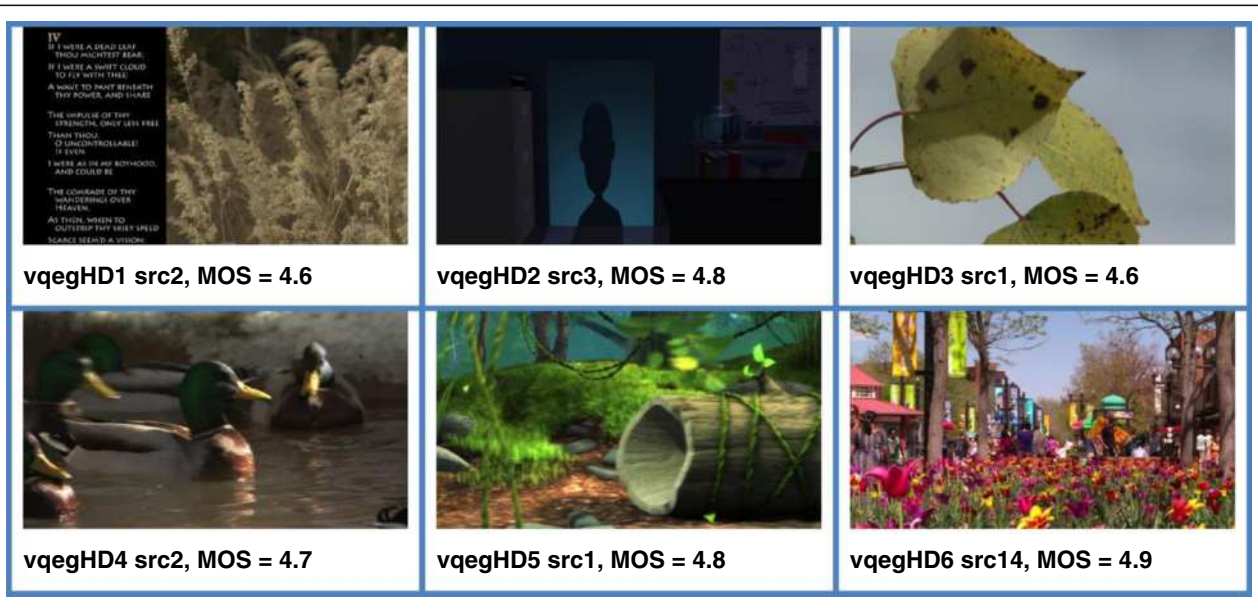

Figure 1 Original sequences with high MOS.

Figure 2 Original video sequences with low MOS.

a large cumulative impact on quality. These differences can be best understood by watching the videos on the Consumer Digital Video Library (CDVL; www.cdvl.org). Sequence 'vqegHD6 src 14' from Figure 1 can be found by searching for the title 'Common SRC 14,' and 'vqegHD6 src 4' from Figure 2 can be found by searching for the title 'NTIA Green Bird.'

Niu and Liu [3] analyze the differences between professionally produced video and amateur video. They propose an algorithm that detects whether or not a video has professional production quality. Niu and Liu describe the quality impact of camera motion, shot length (e.g., duration between scene cuts), lighting, color palette, noise, focus, and depth of field.

Choose scenes that evenly span a wide range of coding difficulty

Video encoders, video decoders, error concealment software, and video quality metrics often adapt to the coding difficulty of the video. Thus, some algorithmic deficiencies appear only in hard-to-code scenes, while others appear only in easy-to-code scenes. If the scene pool for a subjective experiment considers only easy-to-code scenes (or only hard-to-code scenes), then the system under test will not be fully characterized. For example, when scenes are coded at a low bitrate, the motion caused by the *I*-frame update is typically difficult to detect in a hard-to-code scene yet becomes obvious in an easy-to-code sequence with very little motion, such as 'NTIA Snow Mountain' from CDVL. As another example, a sequence with a person running across the scene can be problematic for some error concealment algorithms because filling in the missing video with prior content causes the running person to disappear.

Easy-to-code scenes are widely available because they are easy to shoot. Finding hard-to-code content is more challenging. To simplify the task of judging scene complexity, we use an objective complexity metric such as

- Spatial perceptual information (SI) from ITU-T Rec. P.910 [4]

$$\mathrm{SI} = \max_{\mathrm{time}}\left\{\mathrm{std}_{\mathrm{space}}[\mathrm{sobel}(I_n)]\right\} \quad (1)$$

- Temporal perceptual information (TI) from ITU-T Rec. P.910 [4]

$$\mathrm{TI} = \max_{\mathrm{time}}\left\{\mathrm{std}_{\mathrm{space}}[I_n - I_{n-1}]\right\} \quad (2)$$

- Criticality from Fenimore et al. [5]

$$\mathrm{SI}(I_n) = \mathrm{rms}_{\mathrm{space}}[\mathrm{sobel}(I_n)] \quad (3)$$

$$\mathrm{TI}(I_n) = \mathrm{rms}[I_n - I_{n-1}] \quad (4)$$

$$\mathrm{Criticality} = \log_{10}\{\mathrm{mean}_{\mathrm{time}}[\mathrm{SI}(I_n) \times \mathrm{TI}(I_n)]\}, \quad (5)$$

where I_n is the luma plane of source sequence image number n, $\max_{\mathrm{time}}$ and $\mathrm{mean}_{\mathrm{time}}$ are the maximum and mean values in the time series, respectively, and $\mathrm{std}_{\mathrm{space}}$ and $\mathrm{rms}_{\mathrm{space}}$ are the standard deviation and root mean square over all pixels in one image, respectively.

Alternatively, the scenes can be classified visually. To estimate coding difficulty, we encode all video content at

a constant, low bitrate. The hard-to-code sequences will have a much lower resultant quality than the easy-to-code sequences.

At a minimum, we recommend the following:

- Two clips that are very difficult to code (e.g., criticality ≥ 3.5)
- Two clips that are very easy to code (e.g., criticality ≤ 2.5)
- One high spatial detail clip (e.g., many small objects, SI ≥ 200 [4])
- One high motion clip (e.g., an object that moves across the screen in 1 s, which corresponds to an angular speed of 33°/s on a Full-HD display at 3H distance)

Consider frequency and placement of scene cuts

Scene cuts interact in interesting ways with quality perception and with video codecs. Winkler [6] analyzes prior research on spatial masking and temporal masking. Scene cuts mask impairments that occur temporally one frame to a few hundred milliseconds after a scene cut (i.e., 'forward masking'). 'Backward masking' can also occur, masking impairments before a scene cut.

Encoders can introduce a new group of pictures in response to a scene cut. This affects the bitrate allocation during encoding and the propagation of transmission errors during decoding. Scene cuts occur very frequently in movies and broadcast television [3]; they do not typically occur in other applications such as videoconferencing or surveillance.

Scene cuts complicate subjective testing. The concern is that the encoded quality may be dramatically different before and after the scene cut due to changes in properties of the video content. The task of judging quality, usability or experience thus becomes more difficult, because the perceived quality changes. Some researchers only select content that does not have scene cuts. This was the prevalent opinion expressed in VQEG and ATIS throughout the 1990s.

The problem is that these results may not fully represent user experiences. This is the prevalent opinion expressed in VQEG today. Our preference regarding scene cuts is to select the following:

- About half of the clips with scene cuts
- One clip with rapid scene cuts (e.g., every 1 to 2 s)
- About half of the clips without scene cuts

Note that the 'differing quality' phenomenon is not unique to scenes with scene cuts. This also occurs spatially or temporally in continuously filmed content. Different parts may be better focused or intentionally blurred, relatively still, or containing significant motion. Any of these variations will trigger quality differences that might make the subject's task more difficult.

Select scenes with unusual properties

We learn the most from unique scenes with extraordinary features that may stimulate anomalous behavior in the transmission chain. For example, consider a scene showing a closeup view of a person. Test subjects know how people should look, move, and sound. Their internal reference helps them notice the unnatural motion of a reduced frame rate. That reduced frame rate may be less obvious when watching a video of a machine. A frame rate or frame freeze will become imperceptible if it occurs during a still or nearly still segment of a video sequence, as would a frame freeze. By contrast, a frame freeze that occurs in the middle of a camera pan will be obvious.

The following scene traits can interact in unique ways with a codec or a person's perception. Our ideal scene pool includes all of these traits:

- Animation, graphic overlays, and scrolling text
- Repetitious or indistinguishable fine detail (e.g., gravel, grass, hair, rug, pinstripes)
- Sharp black/white edges
- Blurred background, with an in-focus foreground
- Night or dimly lit scene
- Ramped color (e.g., sunset)
- Water, fire, or smoke (for unusual shapes and shifting patterns)
- Jiggling or bouncing picture (e.g., handheld camera)
- Flashing lights or other extremely fast events
- Action in a small portion of the total picture
- Colorful scene
- Small amounts of analog noise (e.g., camera gain from dim lighting)
- Multiple objects moving in a random, unpredictable manner
- Visually simple imagery (e.g., black birds flying across a blue sky)
- Very saturated colors
- Rotational movement (e.g., a carousal or merry-go-round seen from above)
- Camera pans
- Camera zoom
- Tilted camera

Consider interlace issues

Interlacing and deinterlacing artifacts can occur when a scene contains edges that move within the frame. Moving diagonal edges are particularly noticeable. These artifacts become particularly visible, and thus objectionable, on moving diagonal edges. The traditional deinterlacing detection sequence is a Silicon Optix test

disc[a] sequence showing an American flag waving in the breeze. Deinterlacing artifacts are easily visible on the high-contrast edges of the red and white stripes. The pictures in the work of Jung et al. [7] and Koo et al. [8] show the impact of deinterlacing problems using this flag sequence. Although any content with moving diagonal edges may be used, interlacing problems are more easily seen on strong contrast edges, which may trigger additional impairments.

New issues when selecting 3D sequences

Content selection in 3D subjective assessment tests is made even more complicated due to the introduction of binocular disparity. To ensure viewing comfort for any given target display size, new restrictions must be applied when shooting or selecting 3D content [9].

Besides the comfortable viewing criterion for 3D content, the inclusion of the stereopsis in the content allows for particular content that is perceived differently from its two-dimensional (2D) counterpart. Win et al. [10] demonstrate that the 2D and 3D versions of the same sequence receive different subjective scores and that simply wearing 3D glasses has a negative impact upon the perceived quality. Jumisko-Pyykkö and Utriainen [11] found that 2D and 3D presentation modes on a small mobile device yielded very different response levels to questions about satisfaction, quality, and acceptability. When asked about their viewing experience, subjects selected very different qualitative terms. Therefore, 3D subjective experiments should include sequences that exercise stereopsis in a variety of ways.

Choose content that avoids visual discomfort and eye fatigue

Because poorly produced 3D content may cause visual discomfort and eye fatigue, content editing and camerawork become a critical factor when selecting 3D content. Professional stereographers seek to maintain the 3D effect while minimizing discomfort. The conventional advice includes restricting the scene depth, positioning important objects in the plane of the monitor, and limiting crossed parallax (i.e., preventing objects from appearing too close to the viewer) [12]. Mendiburu [13] provides an in-depth primer on 3D camerawork from a stereographer's perspective. 3D@Home (www.3dathome.org) provides a variety of useful information, including a tutorial on adapting 3D content to different display technologies, recommendations for creating 3D content, and a list of training courses.

Several research studies have examined this issue from an engineer's perspective. Lee et al. [14] analyze the underlying properties of reconstruction of 3D content on stereoscopic screens and the allowable depth budget. Chen et al. [15] propose a detailed shooting rule for 3D content, based on the results of several subjective experiments.

Include a variety of motion directions

One commonly used advantage of 3D is the ability to portray motion in the depth plane (e.g., motion toward or away from the viewer). This effect is often used in production. However, fast depth motion can lead to visual discomfort [16,17]. The non-translational motion behavior (i.e., motion in the depth plane) challenges video coding algorithms, causing the appearance of different artifacts in the two views, which often results in binocular rivalry.

Fast motion was already mentioned for 2D sequences (e.g., include one high motion clip). Fast planar motion (i.e., motion parallel to the screen) becomes an important feature for 3D, because it can introduce visual discomfort in 3D viewing [18,19]. The magnitude of visual discomfort caused by objects moving in the depth plane depends on the position and size of the object, as well as its motion amplitude and speed [20].

The pop-out effect occurs when an object with a large positive disparity is shown in front of the screen for a limited duration. The pop-out effect stresses the 3D effect and is therefore often used as a short-term attractor for 3D visualization.

A 3D scene pool should include at least two sequences with slow motion in depth, sequences with different amounts of planar motion (e.g., slow to fast), and one pop-out effect. The goal is to maximize motion diversity while avoiding motion that causes visual discomfort. The professional stereographer filming guidelines mentioned above can help the experimenter make this assessment, but, in the end, the experimenter is responsible for conducting an ethical experiment that will not cause subjects undue discomfort. The level of visual discomfort caused by candidate 3D sequence should be subjectively assessed by a panel of experimenters during 3D sequence selection.

Vary the depth budget and disparity

Depth budget is a term that combines the positive and negative parallax into a single measurement [12]. The depth budget determines the nearest and furthest objects that the viewer perceives using stereopsis. As the depth budget increases, the differences between what the left eye and right eye see increase.

A large depth budget is caused by shooting 3D sequences at short distances (e.g., filming objects less than 5 m away) or using hyper-stereoscopic shooting (usually resulting from separating the two cameras at a large distance compared to their zoom factor). Sequences with a large depth budget are susceptible to transmission degradations. For example, their coding gain due to

redundancy reduction between the views may be limited by the large disparity and the number of occlusion regions. A large depth budget can also cause crosstalk artifacts to become more pronounced.

A small depth budget is caused by shooting 3D sequences at greater distances with a higher zoom factor but without increasing the interocular distance (e.g., filming objects more than 6 m away) or limiting the depth information (i.e., distance between objects from the camera's point of view). This reduces the perceived depth effect, with the limiting case being 2D. A small depth budget decreases the added value of 3D but generally also reduces visual degradations due to transmission and display properties. A small depth budget minimizes viewer discomfort.

Another interesting disparity effect occurs when the object of interest is not the closest object to the camera. Such sequences can cause unexpected perception issues, because visual attention is attracted by close objects [21]. Our ideal 3D scene pool includes a variety of depth budgets and focal objects at differing disparities.

Look for scenes that interact in unique ways with 3D

The stereoscopic appearance of graphical animations or cartoons differs significantly from natural content. First, optimal camera positions for reconstruction of virtual 3D scenes may be guaranteed. Second, cartoon-type content often contains high image contrasts between flat textured regions; this does not occur in natural content. Likewise, coding algorithms respond differently to mixed content sequences with pronounced contours such as vector graphics.

We vary the distribution of small structures and large structures, fine details, and uniform areas. This is known as frequency distribution, and it influences the perception of 3D [22].

Subtitles or other graphical overlays have large occlusion areas and a high contrast between the foreground text and the video background. Subtitles and other graphical overlays may be particularly impacted by coding and transmission algorithms when contours get smoothed. The depth position of the foreground text may become less obvious, and depth cue conflicts may occur with the background.

Experimental design and implementation issues

The goal of experimental design is to objectively answer a question about an opinion and reach statistically significant conclusions. The nature of human perception inherently confounds all of the variables involved, which adds difficulty.

The critical issue here is that scene choices do not bias the results - either by indicating differences where none exist or by missing significant differences in the response of a variable. When these errors stem from the choice of scenes, the error likely cannot be detected unless an additional subjective test is conducted.

Choose a rating method that solves editing and camerawork problems

Some subjective scales may reduce the impact of editing and camerawork on the ratings. We recommend the following subjective scales for experiments where there is a need to minimize the impact of editing and camerawork:

1. Double stimulus impairment scale (DSIS), also known as degradation category rating, ITU-T Rec. P.910
2. Pair comparison (PC), also known as double stimulus comparison scale and comparison category rating, ITU-R Rec. BT.500
3. Double stimulus continuous quality scale (DSCQS), ITU-R Rec. BT.500
4. Subjective assessment of multimedia video quality (SAMVIQ), ITU-R Rec. BT.1788
5. Simultaneous double stimulus for continuous evaluation (SDSCE), ITU-R Rec. BT.500

These subjective rating scales reduce the impact of editing and camerawork by exposing subjects to both the impaired and the source video, using the source video as a point of reference. The differences between rating methods can be best understood by examining the treatment of the reference video and the comparison task performed by the subject:

- *Labeled reference*. Subjects are told that they are observing the source video.
- *Unlabeled reference*. Subjects are unaware that they are observing the source video.
- *Direct comparison*. Subjects watch two versions of the same sequence and then rate the perceptual difference.
- *Implied comparison*. Subjects watch two or more versions of the same sequence and then rate the quality of each sequence on the same scale. The subjects do not explicitly rate the perceptual difference themselves, yet it is assumed when looking at the rating interface that this comparison will be made by the experimenter.
- *Indirect comparison*. Subjects watch and rate each sequence separately. The unlabeled reference sequence is hidden among the sequences to be rated. The source and impaired sequence ratings are subtracted to compute a difference rating, but the subjects would typically not know that this will occur.

For implied and indirect comparisons, the ratings from the source and impaired sequences are subtracted to compute a difference rating. We expect the standard deviation of these difference ratings to be slightly larger than the standard deviation of the ratings of the individual sequences. The actual impact on the standard deviation will depend upon how much correlation exists between the ratings of the source sequence and the ratings of the impaired sequence. If uncorrelated, we would expect an increase of 41% (i.e., a square root of two increases in the standard deviation). If perfectly correlated, the standard deviation of the difference will be 0. Data from the VQEG reduced reference and no reference (RRNR-TV) test show an 18.1% increase for the 525-line experiment and a 9.6% for the 625-line experiment. These tests were conducted using the ACR with hidden reference (ACR-HR) method from ITU-T Rec. P.910.

We do not recommend the use of indirect comparisons when trying to minimize the impact of editing and camerawork on ratings. An indirect comparison is used by the ACR-HR method. The same technique can be used with single stimulus continuous quality evaluation from ITU-T Rec. BT.500; however, that technique has not been standardized.

Table 1 identifies the approach used by each of the recommended subjective methods. As of this paper's publication, no studies have been published demonstrating an in-depth analysis of how effective these methods are at eliminating the impact of production quality on subjective ratings.

Do not skimp on your scene total

Experimental design is always a compromise between the number of impairments, the number of scenes, and each subject's participation time. It is tempting to reduce the number of scenes (or subjects) so that the number of impairments can be increased. The problem is that only limited conclusions can be drawn when the degradations are analyzed over a narrow range of contents, and those conclusions may not generalize. This makes it impossible to accurately characterize a system under test. The following guidelines optimize the ratio of content variety for the number of degradations.

An entertainment-oriented subjective video quality test of 2D content should use a scene pool containing approximately eight to ten clips. A robust 3D subjective test requires 10 to 14 sequences, because of the additional factors introduced (e.g., motion direction, depth of field, disparity, and unique 3D content interactions).

Table 1 Strategies for reducing the impact of editing and camerawork on subjective ratings

	Direct comparison	Implied comparison
Labeled reference	DSIS	SAMVIQ, SDSCE
Unlabeled reference	PC	DSCQS

The impact of the number of scenes on an experiment can be seen in Pinson et al. [23]. This article analyzes 13 subjective experiments. Each explored the relationship between the following:

- Audio subjective quality (a)
- Video subjective quality (v)
- The overall audiovisual subjective quality (av)

One way to measure this is the Pearson correlation between av and the cross term ($a \times v$). Figure 3 shows a histogram of these correlations, split by the number of scenes in the experiment:

- Limited (one or two)
- Normal (five to ten)

The former spans a range of Pearson correlation from 0.72 to 0.99, indicating that chance played a large role. The latter are tightly clustered, indicating a high degree of repeatability.

Avoid overtraining by maximizing diversity

A common problem is selecting scenes from one small pool of video content. This biases research results toward characteristics of those video sequences. Overtraining is a likely by-product of small scene pools or reusing the same sequences in multiple experiments. For example, an objective model might yield very poor quality estimates when exposed to new content or an encoder might yield very poor quality for some content types (e.g., fire, smoke, confetti, a wood parquet floor, fireworks, saturated red, long dissolves, a pinstripe shirt). Instead, we encourage you to find new sequences for each experiment.

Poor scene selection can invisibly bias experimental results. This is easier to see by examining an experiment with good scene selection. For example, Barkowsky et al. [24] analyze a subjective test that investigated the quantization parameter (QP) parameter on the quality of H.264 encodings. Quantization is the primary algorithmic cause of lost information in the MPEG-2 and H.264 video encoders, so QP is directly linked to image degradation. This experiment's nine scenes were chosen using the criteria described in the section 'Basic scene selection.' Six scenes show similar QP/quality response curves, while the other three show unique behaviors. Without those three scenes, the reported ability of QP to predict quality would have been inflated.

A small number of websites host professionally produced source video content. CDVL [25] is a repository of broadcast quality video content and provides free video clip downloads of video clips for research and

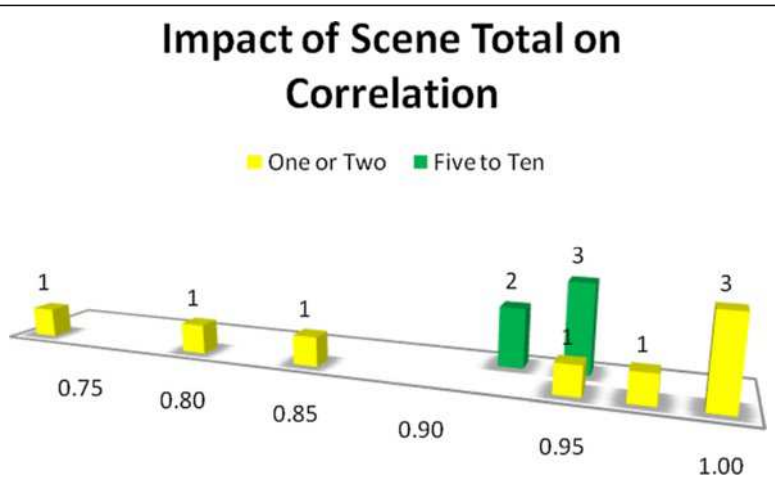

Figure 3 An analysis of 13 subjective tests shows that experiment accuracy depends upon the scene total.

development purposes. CDVL's goal is to foster research and development into consumer video processing and quality measurement. Winkler [26] identifies two other websites that provide uncompressed source video, plus 27 subjective quality image and video databases.

Freely available 3D content is rare. Some 3D sequences are provided by Urvoy et al. [27], Goldmann et al. [28], and Cheng et al. [22].

Be aware of test format implications

The format of a subjective video quality experiment is intentionally artificial. Aural clues that are normally provided by the accompanying audio are missing. Consider these real-world examples:

- The video freezes at the beginning of a movie when a character is introduced, and the character's name is overlaid. The viewer knows that this is intentional because the music and voiceover continue.
- The video flickers quickly between the picture and black to signify that we are seeing past events. A sound effect indicates that this is a flashback.
- For artistic reasons, the video has a digital effect overlay of another color (see Figure 4, left). The edge pattern is similar to what might occur with a transmission error.
- The picture intentionally contains impairments typical of old film, to imply that events occurred long ago (see Figure 4, right).
- Very rapid scene cuts (e.g., 0.25 s apart) can make it appear that the video is fast-forwarding or skipping content in response to network problems.

Without the extra information from the audio track, these artistic effects are indistinguishable from network transmission errors, channel switches, display issues, or other sources of degradation.

Perform scene selection on the device to be tested

Video quality subjective testing has traditionally involved uncompressed video played to broadcast quality monitors. This controlled for the effect of the video playback and monitor from the data and helped us focus on video encoding, network transmission, and video decoding.

Subjective testing on mobile devices must use compressed playback and lower quality monitors - and account for their confounding impact on the subjective data. The computer that used to view, select, edit, and

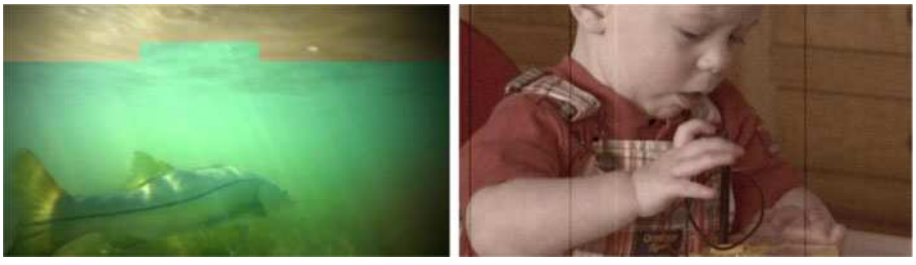

Figure 4 Artistic video effects that may look like impairments: color overlay (left) and old film (right).

impair the video is probably a more powerful computer - perhaps a high-end PC with a large monitor. Switching to the device under test will impact the appearance of your sequence [29]. We recommend that you always perform final scene selection on the device under test.

Semi-automatic selection of source scene pools

Selecting a non-biased, well-balanced source sequence set for a subjective experiment may prove difficult for a single researcher. The following suggests an approach to achieve semi-automatic scene selection.

Begin by enumerating attributes for the available video sequences. This often requires a subjective judgment that should ideally be deferred to a group of observers. Possible attributes include whether or not the scene contains rapid scene cuts, the amount of saturated color presence, and how professional is the editing. Some attributes are concrete and thus Booleans (e.g., whether or not the scene contains flashing lights), but most of the properties are measured along a subjective scale (e.g., to what extent is this a night scene). To reduce selection bias, the definition of each attribute should be established and then a panel of observers asked to rate the attributes of each available video sequence (e.g., 100 videos from the CDVL database).

This approach does not necessarily require standardized environments, and thus, using crowd-sourcing technologies may be appropriate [30]. After obtaining a vector of attributes for each video sequence, data mining algorithms may be applied. Consider the following semi-automatic algorithm:

A. An expert in subjective experiment preparation picks an initial video sequence.
B. The semi-automatic algorithm then suggests a set of video sequences which correlate least to the selected video sequence over all scale value dimensions (see Figure 5).
C. The researcher then picks a second video sequence.

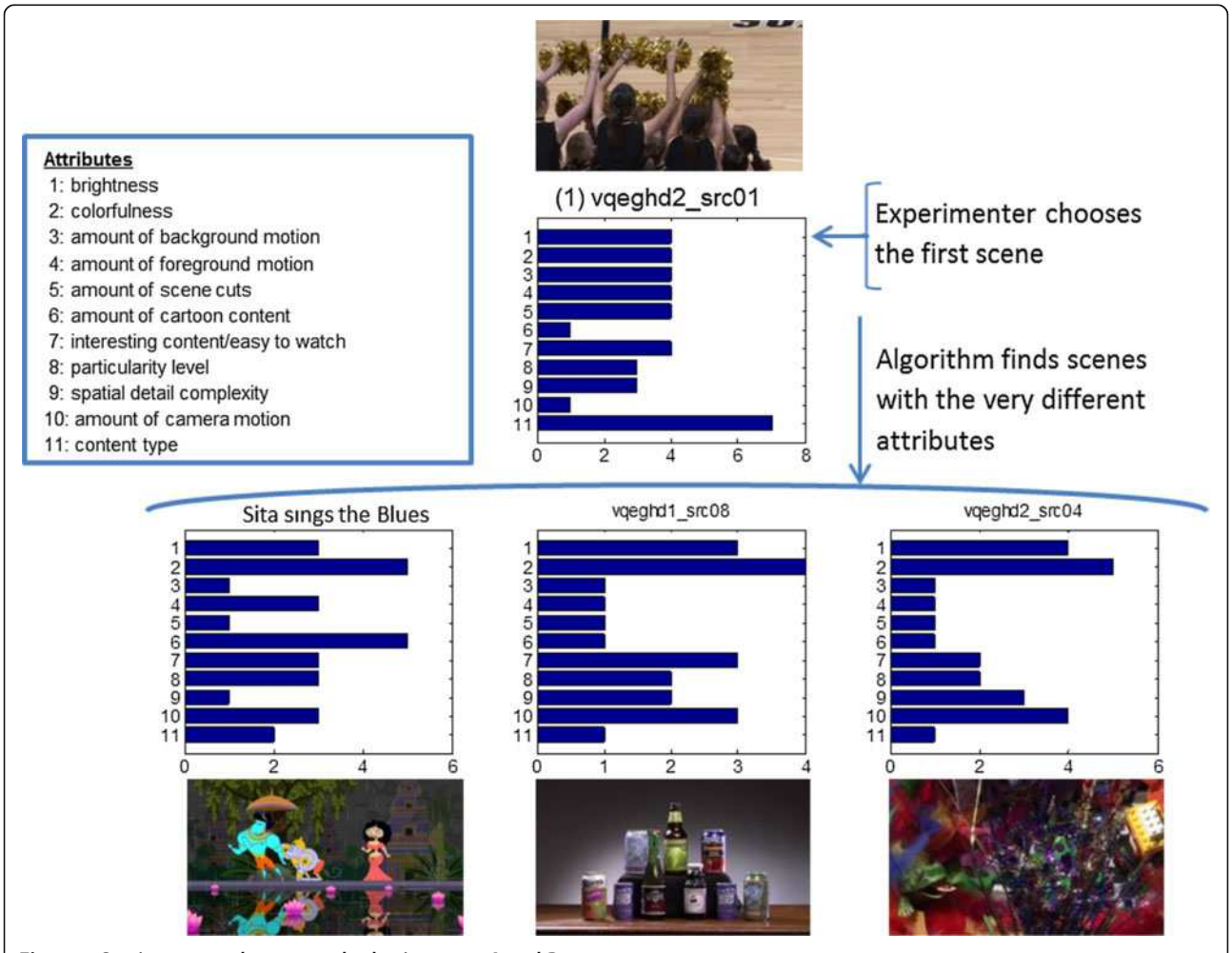

Figure 5 Semi-automated scene pool selection steps A and B.

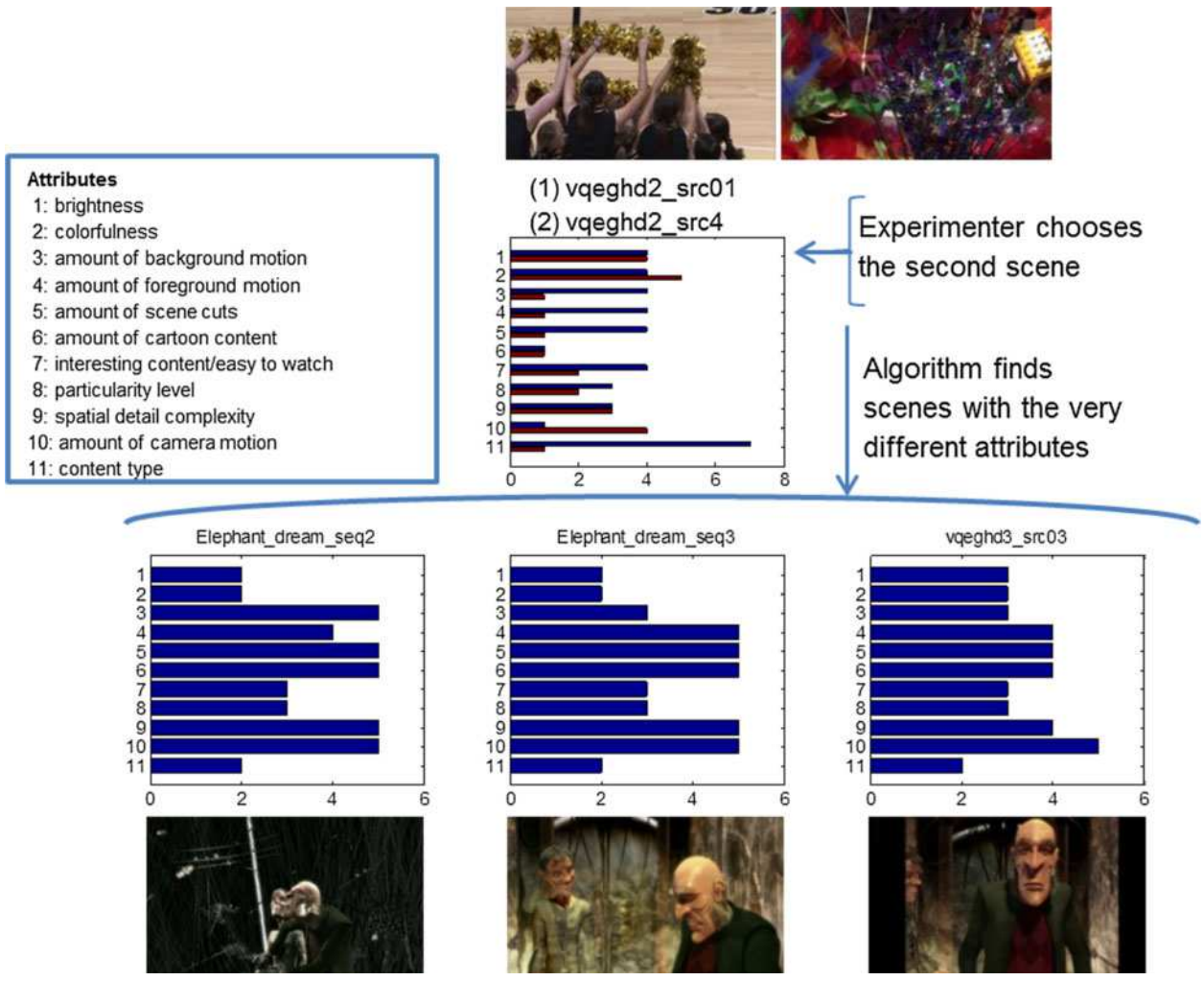

Figure 6 Semi-automated scene pool selection steps C and D.

D. The algorithm takes into consideration the properties of both sequences to recommend options for the third scene (see Figure 6).

This selection and search process continues iteratively (i.e., steps C and D). This allows the selection of video sequences to be more objective and avoids bias while still allowing for interaction when particular video properties are not taken into consideration by the automation.

Figures 5 and 6 illustrate the semi-automatic scene pool selection algorithm using a database of about 200 video sequences and the 11 attributes displayed in Figure 5. In a preliminary experiment, three observers rated each attribute on a scale of one to ten.

The criterion for selecting the most diverging video sequences was calculated as follows. The distance metric was Pearson linear correlation coefficient, calculated over all 11 attributes. The distance was measured between each of the selected sequences and each of the remaining candidate sequences. The distance metric was averaged over the selected sequences (e.g., one sequence for Figure 5 and two sequences for Figure 6). The three sequences with the maximum average distance (i.e., minimum correlation) are shown in Figures 5 and 6. Notice that the algorithm proposed three very different options for the second sequence in Figure 5. However, all three proposals for the third sequence were cartoon sequences with mostly dark features, as may be found in the open-source movie 'Elephants Dream' [31].

Conclusions

The correct selection of scene pools for subjective experiments has been previously mostly limited by content availability. Limited research has been performed on the influence of source variety selection in subjective experiments with respect to reproducibility of assessment results. In most cases, only the influence on degradations, such as coding performance, has been studied.

This paper proposed guidelines for the selection of scene pools with a large variety of content, including a semi-automatic selection process. Mostly 2D video content has been addressed, as it is widely available. For the newly available stereoscopic content dimension, guidelines have been proposed which are meant to facilitate the collection of meaningful source video content or to provide hints for producing or shooting missing content types. Similar guidelines for scene pool selection should be developed for other types of subjective assessments, for example, audiovisual quality assessment.

Endnotes

[a]Certain commercial equipment, materials, and/or programs are identified in this report to specify adequately the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration nor does it imply that the program or equipment identified is necessarily the best available for this application.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors would like to thank technical copy editor Lilli Segre, technical reviewer Wade Allen, technical reviewer Andrew Catellier and editorial sponsor John Vanderau.
Margaret H. Pinson's work was funded by the US government through the NTIA. Patrick Le Callet's and Marcus Barkowsky's work is financed by the French ministry for higher education and research. It was partly conducted within the scope of the PERSEE project which is financed by ANR (project reference: ANR-09-BLAN-0170), and the JEDI (Just Explore DImensions) ITEA2 project which was supported by the French industry ministry through DGCIS.

Author details
[1]National Telecommunications and Information Administration (NTIA), Institute for Telecommunication Sciences (ITS), U.S. Department of Commerce (DOC), 325 Broadway St, Boulder, CO 80305, USA. [2]IRCCyN CNRS UMR 6597, Polytech Nantes, Université de Nantes, LUNAM Université, rue Christian Pauc, BP 50609, Nantes Cedex 3 44306, France.

Received: 29 March 2013 Accepted: 9 August 2013
Published: 28 August 2013

References

1. VQEG, *Report on the validation of video quality models for high definition video content, (Video Quality Experts Group, June 2010)*. http://www.vqeg.org/. Accessed August 2013
2. T Tominaga, T Hayashi, J Okamoto, A Takahashi, *Performance comparisons of subjective quality assessment methods for mobile video. Paper presented at the 2nd Quality of Multimedia Experience (QoMEX)* (Trondheim, 2010)
3. Y Niu, F Liu, What makes a professional video? A computational aesthetics approach. IEEE Trans Circ Syst Video Tech **22**(7) (2012)
4. ITU, *ITU-T Recommendation P.910. Two Criteria for Video Test Scene Selection. Section 6.3* (ITU, Geneva, 1994)
5. C Fenimore, J Libert, S Wolf, *Perceptual effects of noise in digital video compression. Paper presented at the 140th SMPTE technical conference* (Pasadena, 1998)
6. S Winkler, Issues in vision modeling for perceptual video quality assessment. IEEE Signal Process Mag **78**(2) (1999)
7. Y Jung, S Yang, P Yu, An effective de-interlacing technique using two types of motion information. IEEE Trans. Consum. Electron **49**(3) (2003)
8. HI Koo, SH Lee, NI Cho, A new EDI-based deinterlacing algorithm. IEEE Trans. Consum. Electron **53**(4) (2007)
9. W Chen, J Fournier, M Barkowsky, P Le Callet, *New requirements of subjective video quality assessment methodologies for 3DTV. Paper presented at the 5th international workshop on video processing and quality metrics (VPQM)* (Scottsdale, 2010)
10. SL Win, J Caviedes, I Heynderickx, *2D vs. 3D visual quality evaluation: the depth factor. Paper presented at the 7th international workshop on video processing and quality metrics (VPQM)* (Scottsdale, 2013)
11. S Jumisko-Pyykkö, T Utriainen, *User-centered quality of experience: is mobile 3D video good enough in the actual context of use? Paper presented at the 5th international workshop on video processing and quality metrics (VPQM)* (Scottsdale, 2010)
12. D Minoli, *3D Television (3DTV) Technology, Systems, and Deployment* (CRC Press, Florida, 2011)
13. B Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen* (Focal Press, Oxford, 2009)
14. S Lee, YJ Jung, S Hosik, YM Ro, *Subjective assessment of visual discomfort induced by binocular disparity and stimulus width in stereoscopic image. Paper presented at SPIE electronic imaging, stereoscopic displays and applications* (Burlingame, 2013)
15. W Chen, J Fournier, M Barkowsky, P Le Callet, *New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone. Paper presented at SPIE electronic imaging, stereoscopic displays and applications XXII* (San Francisco, 2011)
16. J Li, M Barkowsky, J Wang, P Le Callet, *Study on visual discomfort induced by stimulus movement at fixed depth on stereoscopic displays using shutter glasses. Paper presented at the 17th international conference on digital signal processing (DSP)* (Corfu, 2011)
17. F Speranza, WJ Tam, R Renaud, N Hur, *Effect of disparity and motion on visual comfort of stereoscopic images, in Proceedings of SPIE stereoscopic displays and virtual reality systems XIII, 6055* (San Jose, 2006)
18. J Li, M Barkowsky, P Le Callet, *The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos. Paper presented at the 3rd international workshop on quality of multimedia experience (QoMEX)* (Mechelen, 2011)
19. M Lambooij, M Fortuin, WA IJsselsteijn, I Heynderickx, *Measuring visual discomfort associated with 3D displays. Paper presented at stereoscopic displays and applications XX* (, San Jose, 2009)
20. F Speranza, WJ Tam, R Renaud, N Hur, *Effect of disparity and motion on visual comfort of stereoscopic images. Paper presented at stereoscopic displays and virtual reality systems XIII* (, San Jose, 2006)
21. J Wang, P Le Callet, S Tourancheau, V Ricordel, MP Da Silva, Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli. J Eye Mov Res **5**(5) (2012)
22. E Cheng, P Burton, J Burton, A Joseski, I Burnett, *RMIT3DV: pre-announcement of a creative commons uncompressed HD 3D video database. Paper presented at the 4th international workshop on quality of multimedia experience (QoMEX)* (Yarra Valley, 2012)
23. M Pinson, W Ingram, A Webster, Audiovisual quality components. IEEE Signal Process. Mag **28**(6) (2011)
24. M Barkowsky, M Pinson, R Pèpion, P Le Callet, *Analysis of freely available subjective dataset for HDTV including coding and transmission distortions. Paper presented at the 5th international workshop on video processing and quality metrics for consumer electronics (VPQM)*, 2010
25. M Pinson, S Wolf, N Tripathi, C Koh, *The consumer digital video library. Paper presented at the 5th international workshop on video processing and quality metrics for consumer electronics (VPQM)* (, Scottsdale, 2010)
26. S Winkler, Analysis of public image and video databases for quality assessment. IEEE J Sel Top Signal Process **6**(6) (2012)
27. M Urvoy, M Barkowsky, R Cousseau, Y Koudota, V Ricordel, P Le Callet, J Gutiérrez, N Garcia, *NAMA3DS1-COSPAD1: subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. Paper presented at the 4th international workshop on quality of multimedia experience (QoMEX)* (Yarra Valley, 2012)
28. L Goldmann, F De Simone, T Ebrahimi, *A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. Paper presented at the electronic imaging (EI), 3D image processing (3DIP) and applications* (San Jose, 2010)
29. A Catellier, M Pinson, W Ingram, *A Webster, Impact of mobile devices and usage location on perceived multimedia quality. Paper presented at the 4th*

international workshop on quality of multimedia experience (QoMEX) (Yarra Valley, 2012)
30. C Keimel, J Habigt, C Horch, CK Diepold, *Video quality evaluation in the cloud. Paper presented at the 19th international packet video workshop (PV)* (Munich, 2012)
31. Blender Foundation, The Orange Project, *Elephants Dream in stereoscopic 3D*. http://orange.blender.org. Accessed March 2013

doi:10.1186/1687-5281-2013-50
Cite this article as: Pinson *et al.*: Selecting scenes for 2D and 3D subjective video quality tests. *EURASIP Journal on Image and Video Processing* 2013 **2013**:50.